DOCUMENT RESUME

ED 237 534 TM 830 784

AUTHOR Hunter, John E.
TITLE Fairness of the General Aptitude Test Battery: Ability Differences and Their Impact on Minority Hiring Rates. Uses Test Research Report No. 46.
INSTITUTION California State Dept. of Employment Development, Sacramento.
SPONS AGENCY Employment and Training Administration (DOL), Washington, D.C.
PUB DATE 83
NOTE 33p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Adults; American Indians; *Aptitude Tests; Asian Americans; Blacks; Comparative Testing; Hispanic Americans; Job Performance; *Minority Groups; *Personnel Selection; Predictor Variables; Psychological Testing; *Psychomotor Skills; *Test Bias; Test Norms; Test Results
IDENTIFIERS *General Aptitude Test Battery

ABSTRACT

This paper reviews the now massive general literature showing that psychological tests are fair to minorities. This literature shows that there is no single group validity, there is no differential validity, and tests overpredict rather than underpredict minority job performance. Further evidence in regard to blacks is introduced from 51 validation studies done by the United States Employment Service. General Aptitude Test Battery norms for Blacks, Indians, Mexican Americans, Orientals, and the majority are compared. Although the majority is higher on cognitive abilities, three out of four minority groups are higher than the majority on psychomotor ability. Thus, there is a varied pattern of rank orders among groups across jobs of different complexity. In particular, it is shown that for jobs of low complexity, the addition of psychomotor ability as a predictor simultaneously reduces adverse impact while increasing the validity and, hence, economic benefits of the use of tests for selection. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED237534

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

USES TEST RESEARCH REPORT NO. 46

FAIRNESS OF THE GENERAL APTITUDE TEST BATTERY: ABILITY DIFFERENCES AND THEIR IMPACT ON MINORITY HIRING RATES

DIVISION OF COUNSELING AND TEST DEVELOPMENT
EMPLOYMENT AND TRAINING ADMINISTRATION
U.S. DEPARTMENT OF LABOR
WASHINGTON, D.C. 20213
1983

TM 830 784

TABLE OF CONTENTS

	PAGE
List of Illustrations	i
List of Tables	ii
Acknowledgement	iii
Abstract	iv
Overview: Test Fairness Versus Racial Imbalance	1
The Scientific Proof that Tests are Fair to Minorities	4
Fairness of the GATB	12
Racial and Ethnic Differences on the GATB	15
Reference Notes	23
References	24

LIST OF ILLUSTRATIONS

Figure Number	Page	
1	7	A hypothetical scatterplot showing the relationship between test scores and job performance scores for some validity study.
2	8	The hypothesized relationship between ability and performance for majority and minority applicants if both groups all have the identical score of 100 on the test but the test is unfair to minorities.
3	10	The empirical disconfirmation of the hypothesis that cognitive tests are unfair to members of minority groups.

LIST OF TABLES

Table Number	Page	
1	13	The extent of overprediction of black job performance by the GATB.
2	17	Differences in aptitude and ability distributions on the GATB.
3	18	Racial and ethnic group means on the GATB general abilities expressed in majority group standard scores.
4	19	Job families determined by complexity.
5	21	Percentage of each ethnic or racial group selected as a function of selection ratio and job complexity.

ACKNOWLEDGEMENT

The United States Employment Service (USES) conducts a test research program for developing testing tools useful in vocational counseling and placement.

The purpose of this series of reports is to provide results of significant test research projects as they are completed. These reports will be of interest to users of USES tests and to test research personnel in state agencies and other organizations.

This paper cumulates the findings of validity studies done by hundreds of analysts working for the U.S. Employment Service over a 45-year span. Special thanks go to John Hawk of the U.S. Employment Service in Washington, D.C., and Ron Boese of the North Carolina Employment Security Commission for the computer analyses they conducted.

This report was written by Dr. John E. Hunter, Michigan State University, under contract to the California Test Development Field Center, California Employment Development Department, Los Angeles, California. The report was prepared for printing by staff of the Western Test Development Field Center, Utah Department of Employment Security.

FAIRNESS OF THE GENERAL APTITUDE TEST BATTERY: ABILITY DIFFERENCES AND THEIR IMPACT ON MINORITY HIRING RATES

ABSTRACT

This paper reviews the now massive general literature showing that psychological tests are fair to minorities. This literature shows that there is no single group validity, there is no differential validity, and tests overpredict rather than underpredict minority job performance. Further evidence in regard to blacks is introduced from 51 validation studies done by the U.S. Employment Service. GATB norms for blacks, Indians, Mexican Americans, Orientals, and the majority are compared. Although the majority is higher on cognitive abilities, three out of four minority groups are higher than the majority on psychomotor ability. Thus, there is a varied pattern of rank orders among groups across jobs of different complexity. In particular, it is shown that for jobs of low complexity, the addition of psychomotor ability as a predictor simultaneously reduces adverse impact while increasing the validity and hence economic benefits of the use of tests for selection.

OVERVIEW: TEST FAIRNESS VERSUS RACIAL IMBALANCE

Any discussion of the racial or ethnic impact of testing must sharply distinguish between two questions: (1) Is the test fair to minority group members? and (2) Does the use of the test produce a racially or ethnically unbalanced work force? To say that a test is fair is to say that the score on the test is an accurate estimate of the person's ability for all groups. Those who hypothesize that tests are unfair to minorities believe that test scores underestimate the ability of minority members. This is a scientific question which can be answered empirically. The first section of this paper will review the now massive amount of data gathered on this question. This review shows the scientific proof that tests are fair to minority groups.

The second section reviews similar data gathered by the U.S. Employment Service. Since the early seventies, the Employment Service has identified workers by race and ethnicity in their validation studies. There are now 51 studies with enough black workers to permit separate analysis. These studies show no differential validity and show that black job performance is over-predicted rather than underpredicted by the General Aptitude Test Battery (GATB). Thus, the tests used by the U.S. Employment Service are fair to minority applicants.

To say that test use produces an unbalanced work force is to say that selection based on the test, results in hiring a smaller percentage of minority applicants than majority applicants. The extent of disparity in hiring rates depends on the difference between group means on test scores. These differences vary drastically from test to test. For example, U.S. Employment Service data (USES, 1970, p. 281) show that Mexican Americans have a mean on cognitive ability (intelligence, verbal ability, numerical ability, etc.) that is about one-half standard deviation below the majority mean. Thus, for a high-complexity job where optimal test use calls for cognitive ability, an employer selecting the top half of majority applicants would hire only the top 31 percent of the Mexican-American applicants. However, the mean for Mexican Americans on psychomotor ability is .18 standard deviations higher than that for the majority. On low-complexity jobs where psychomotor ability would be a key predictor, an employer hiring the top half of majority applicants would hire 57 percent of the Mexican-American applicants. Thus, the same employer could have an imbalance in hiring rates going in opposite directions for different jobs.

Multiple regression shows that high-paying, white-collar jobs can be best predicted using a cognitive ability composite score (Ghiselli, 1973; Pearlman, Schmidt, and Hunter, 1980; Schmidt, Hunter Pearlman, and Shane, 1979; Hunter, Note 3). Thus, selection of an optimally productive work force would mean underrepresentation of blacks, Mexican Americans, and Indians. Any scheme which uses tests in such a way as to reduce the ethnic imbalance in hiring rates necessarily reduces the average productivity of the

applicants hired. If tests were unfair to minority applicants, then it would be possible to create new tests to resolve the problem. But tests are fair, and the differences in ability between ethnic groups are real. If these differences stem from cultural disadvantage, then the differences may disappear over the next several generations. However, at the present time, there is no way to avoid the trade-off between high productivity and ethnic imbalance in hiring.

The third section of this paper will present the data on racial and ethnic differences for the GATB used by the U.S. Employment Service. There are large differences between groups on cognitive ability, moderate differences on perceptual ability, and much smaller differences on psychomotor ability. In fact, three out of four minority groups actually have higher means than the majority on psychomotor ability.

Since different jobs depend on different abilities, the impact of optimal test use on hiring rates varies across jobs. For high-complexity jobs, the key ability is cognitive ability and only Orientals have hiring rates near to those for the majority. However, for jobs of low complexity, the hiring rates for minority groups equal and even exceed those for the majority. Only blacks have a lower hiring rate than the majority for jobs of the lowest complexity.

The GATB is unique in finding jobs where minority groups would be hired at rates higher than the majority. This follows from the fact that the GATB is the only major battery using psychomotor ability to predict job performance. For jobs of low complexity, the addition to psychomotor ability as a predictor simultaneously reduces adverse impact while it increases validity. Thus, for jobs of low complexity, the addition of psychomotor ability improves economic benefits of testing at the same time that it reduces ethnic imbalance in hiring rates.

The choice between high productivity and ethnic balance can be treated as an ethical decision. Hunter and Schmidt (1976) showed that so-called "statistical modes of test fairness" discussed in the professional literature are actually different ways of introducing quotas into hiring.

However, the ethical discussion of test "fairness" mostly ignored the economic costs that result from nonoptimal use of tests. If the employer has increased labor costs because of the use of quotas of some sort, then these costs must be passed on. Manufacturers pass the costs on in the form of higher prices. This results in lower sales and hence in lower employment; especially in cases where the American firm is in direct competition with foreign manufactures. This increase in unemployment hits hardest among minority workers. Thus, jobs gained by some minority workers are lost for other minority workers; and the economy as a whole suffers severely.

The situation is even more dramatically complicated in the public sector. Hunter (Note 1) estimated that abandonment of a cognitive ability test for the selection of police officers in Philadelphia would result in increased labor costs of $180 million over a 10-year period. Philadelphia cannot further increase its business taxes without a massive exodus of businesses and hence employers to the suburbs. Thus, the $180 million can only be paid in one of two ways: (1) reduction in the quality of police protection or (2) reduction in city services in other areas. Police protection is most important in the high-crime areas. High-crime areas are heavily overrepresented by minority citizens. Thus, the ultimate cost of reduced police protection is borne largely by minority citizens. Reduction in city services in other areas would mean reduction in social services such as subsidized medical services. Again, these services are most heavily used by minority citizens. Thus, the ultimate cost of reduced city services would be borne largely by minority citizens. That is, ethnic balance in police hiring results in little economic benefit to minority citizens while resulting in reduced city services for all.

The economic costs of various schemes for achieving ethnic balance can be calculated. Hunter, Schmidt, and Kauschenberger (1977) analyzed utility differences for the various statistical models of fair test use. These results were recently replicated and extended by Cronbach, Yalow, and Schaeffer (1980). Under most conditions, the professionally derived procedures for achieving ethnic balance result in a loss of economic benefits of 15 percent or less.

However, the Equal Economic Opportunity Commission has been persuading employers to the use of a radically different method of achieving ethnic balance: the low-cutoff method. In this procedure, the test is used only to screen out the extremely poor prospects, usually the bottom 10 to 20 percent. Applicants are then hired randomly from among those above this very low-cutoff score. Hunter (Note 1, Note 4) and Mack, Schmidt, and Hunter (Note 5) have shown the cost of the low-cutoff procedure to be disastrous. At least 85 percent of the reduction in labor costs is lost by use of the low-cutoff method. This 85-percent loss is in comparison to a 15-percent loss resulting from use of population quotas in a comparable situation.

But the low-cutoff procedure is not only a disaster economically, it does not even achieve its racial and ethnic aims. The number of minority members hired is higher for quotas than for the low-cutoff method. Thus, by any criterion, the low-cutoff method is somewhat worse for minority applicants and disasterously worse for employers (and for those who pay the ultimate bill).

There are additional problems created by any kind of quota hiring procedure which stem from the fact that the increase in hiring among minorities is achieved by hiring a subset of minority workers whose ability level is below that for all other workers hired. First, low-ability workers will on the

average be the low-performance workers. If the organization fires low-performance workers, then the quotas are undone. If the poor-performance workers are kept on, then they become known as "affirmative action" workers. This is demoralizing for the poor worker and creates considerable resentment among the better workers who view themselves as "carrying" the poor worker. There is also great cost for the higher performance black worker since there is no recognition that the high-ability black worker is there on merit.

Second, if higher positions within the organization are filled from within, then using lowered standards for minority workers at the entry level produces severe problems for later promotion. If promotion is done on the basis of high performance on the entry-level job, then merit promotion quickly produces an entry-level worker population which is heavily weighted with low-performance, low-ability workers. Thus, new, high-ability hires (largely white) will be promoted ahead of low-performance workers who have been on the job for some time. This may appear to be even more unfair than using merit hiring for the entry-level job in the first place.

Third, the promotion problem is even worse if the higher order jobs require more cognitive ability than the lower level job. Now maximum validity promotion would call for the use of tests as well as performance as a basis for promotion. The merit-versus-quotas dilemma is then moved en mass from entry-level hiring to promotion.

THE SCIENTIFIC PROOF THAT TESTS ARE FAIR TO MINORITIES

Overview

Fifteen years ago, industrial psychologists became generally aware of the large difference between blacks and whites in mean cognitive ability. Since most psychologists believed at that time that there could be no real difference between racial groups in cognitive ability, many assumed that the differences in test scores might mean that the tests were biased against blacks. The theory behind this hypothesis was this: Tests are developed by middle-class, white psychologists in terms of their own cultural ways of thinking and perceiving. Black culture is so different from white culture that items which have one meaning for white applicants might have a different meaning (or no meaning) to black applicants. Thus, test scores for black applicants would underestimate their actual ability level. Many pointed to the known differences between black and white English dialects as the basis for a linguistic bias in tests written by whites.

Actually, there was plenty of evidence available even 15 years ago to show that the cultural hypothesis is false; though that evidence had not yet been collated. In particular, there are many nonverbal tests of cognitive ability and mean differences between blacks and whites are just as high on the nonverbal as on the verbal tests. However, there was little evidence bearing on this question within the employment area at that time. Since then

hundreds of data sets have been accumulated. These studies show that tests are just as valid for blacks as for whites, and that test scores do not underestimate ability for blacks. Similar evidence has also been accumulated for Hispanic applicants. This evidence will be reviewed below.

The key to testing the hypothesis of test bias is to state the hypothesis in terms which can be assessed empirically. This has been done in three ways. The most extreme form of the test-bias hypothesis is the assertion that black culture is so alien to white culture that a test might be completely meaningless to blacks. Thus, a test which is a valid predictor of job performance in some setting for whites might be completely invalid for blacks. This is known as the hypothesis of "single group validity." A less extreme version of this hypothesis is the assertion that a test will be less meaningful for blacks than for whites. Thus, any given test will be less valid for blacks than for whites. This is the hypothesis of "differential validity." Finally, there is the mildest form of the hypothesis: Some items on a test will be meaningless for blacks. Thus, while the rank order of scores for black applicants is correct, and hence the test is just as valid for blacks as for whites considered separately, the scores for blacks will be systematically lower than the scores for whites of equal ability because the blacks will miss the bias items. This is the hypothesis of "underprediction of black performance." All three hypotheses have been extensively tested and all three have been found false.

Evidence Against the Single Group Validity Hypothesis

If a test is a valid predictor of job performance for whites, then will it also be valid for blacks? At one point there seemed to be evidence to suggest that tests valid for white applicants were sometimes not valid for blacks. However, this evidence has subsequently been shown to be an artifact of the statistical procedure used. Since most studies have data for many more white workers than black workers, a given correlation is much more likely to be statistically significant for whites than for blacks. Thus, separate significance tests do not give an accurate assessment of the single group validity hypothesis.

The first statistically correct study in this area was done by Schmidt, Berner, and Hunter (1973). A number of studies had reported significant correlations for whites but not for blacks (apparent single group validity). However, they noted that in each such study there was a vast disparity in the sample sizes for the two racial subgroups. Thus, the same sized correlation would be significant for whites but not for blacks. They devised a procedure for cumulating evidence across studies in such a way that would control for differences in sample size. They applied this cumulative analysis to 410 sets of validity data, 249 studies using supervisor ratings as the job performance measure, and 161 studies in which a job sample test or production record was used to measure job performance. This cumulative analysis showed that findings of single group validity are entirely an artifact of differential sample size. In fact, their cumulative analysis suggested that there

are no differences in validity between blacks and whites at all. The Schmidt, Berner, and Hunter single group validity cumulation has been subsequently replicated three times (O'Connor, Wexley, and Alexander, 1975; Boehm, 1977; and Katzell and Dyer, 1977).

Evidence Disconfirming the Differential Validity Hypothesis

The Schmidt, Berner, and Hunter (1973) procedure does not have great statistical power against the more gentle hypothesis that tests are less valid for blacks than for whites. Thus, while they did show that there is no single group validity, they did not conclusively show that there is no differential validity. Hymphreys (1973) also showed that doing separate significance tests is an inappropriate way of assessing differences in validity between races. He suggested testing the difference between the correlations for statistical significance. In 1974, the American Psychological Association Standards for tests, expressly endorsed the Hymphreys position.

Katzell and Dyer (1977) and Boehm (1977) claimed to have found evidence for differential validity using a cumulative form of the Hymphreys' procedure. They applied the Hymphreys' test to a cumulation of data across many studies and counted the number of times that they found statistical significance. They found more than 5 percent significant findings and therefore concluded that they had found evidence for differential validity. However, there was a conflict between their findings and the hypothesis of differential validity. When they looked to see which correlation is bigger, both studies found that the validity coefficient for blacks was just as likely to be larger than the coefficient for whites, as smaller. That is, they found validity for blacks to be just as high as validity for whites. The discrepancy between these findings was explained by Hunter and Schmidt (1978) who noted that both studies had preselected the pairs of correlations to be tested. Hunter and Schmidt showed mathematically that this preselection would have the effect of producing a spuriously high (i.e., as much as 20 percent) number of significant differences among the subsamples of considered studies.

In a non-preselected sample of 1,190 pairs of regression lines, Bartlett, Bobko, Mosier, Hannan and (1978) found a chance level 5.21 percent differences in slopes. Hunter, Schmidt and Hunter (1979) have used a variety of more powerful cumulation procedures with the same result; differences in validity between racial groups are the statistical artifact of the use of small sample sizes.

A cumulative study of differential validity for Hispanic workers has been done by Schmidt, Pearlman, and Hunter (1980). They located 1,323 data sets in which test-criterion correlations are given for both majority and Hispanic workers. Their initial analysis showed that 11 percent of the correlations were significantly different and the cumulative chi-square test was statistically significant. However, a further check showed that over half of the

significant differences occurred in one small study with very small sample sizes, namely the Rosenfeld and Thornton (1976) study at Site 4. At this one site, the average validity for the sample of 62 majority workers was a minus .16 while the average for the sample of 49 Hispanic workers was plus .16. That is, most of the significant differences were from one study in which the test was apparently valid for Hispanic workers and not valid for the majority. However, these results for the majority group are highly suspect since they contradict the results found in the other three sites in the Rosenfeld and Thornton (1976) study and contradict a large number of studies done on similar tests in other settings for the same job. If the data for this one suspect study are deleted, then there are 1,128 data sets left. The number of significant differences is less than 6 percent, well within the range of chance expectation.

The results of these cumulative studies are completely clear. The validity of a test in the employment area will be exactly the same for blacks as for whites. The validity of the test will be exactly the same for Hispanics as for majority white workers. Thus, in any given setting, the validity of a given test for predicting job performance will be the same for whites, for blacks, and for Hispanic workers. There is no differential validity.

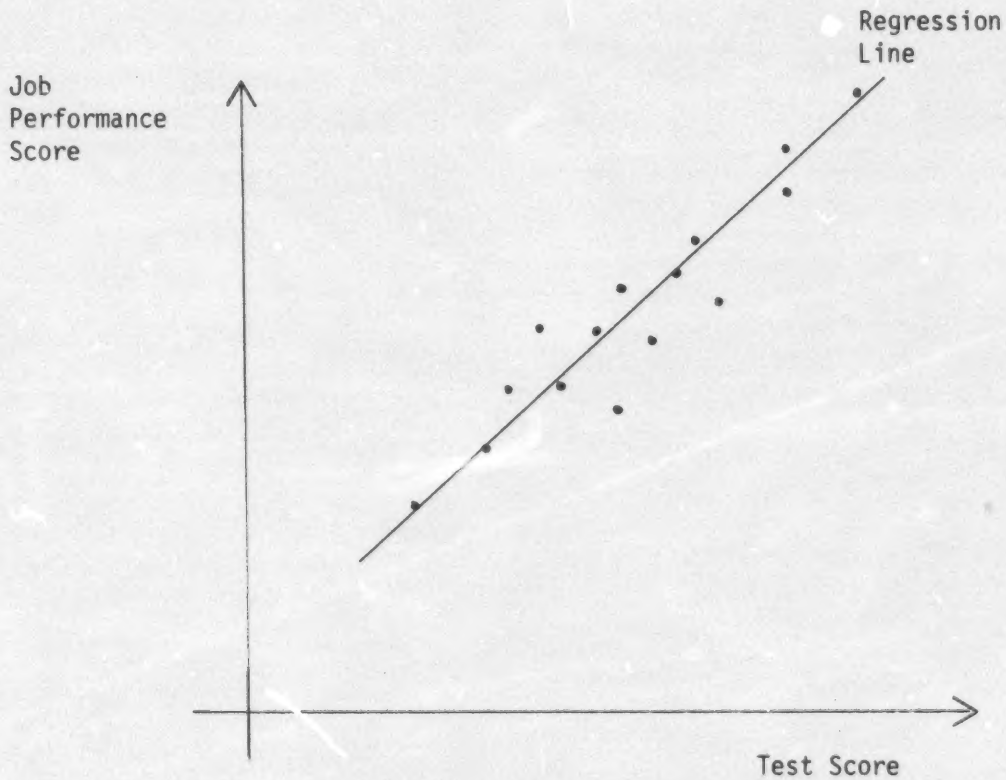


Figure 1. A Hypothetical Scatterplot Showing the Relationship Between Test Scores and Job Performance Scores for Some Validity Study

Evidence Disconfirming the Hypothesis that Tests Underpredict Minority Job Performance

The least extreme form of the hypothesis that tests are unfair to minorities is the claim that only certain items on each test are biased. If only certain items were biased, then the test as a whole would still be as valid for blacks as for whites considered separately. However, the test scores for blacks would be systematically lower than those for whites of the same ability level because blacks would miss the biased items. If it were true that tests underestimate black ability, then it would follow that test scores would underpredict black performance on the job. This in turn leads to the prediction that if tests were biased against blacks, then the regression line for blacks would lie above the regression line for whites. The data show just the reverse of this to be true.

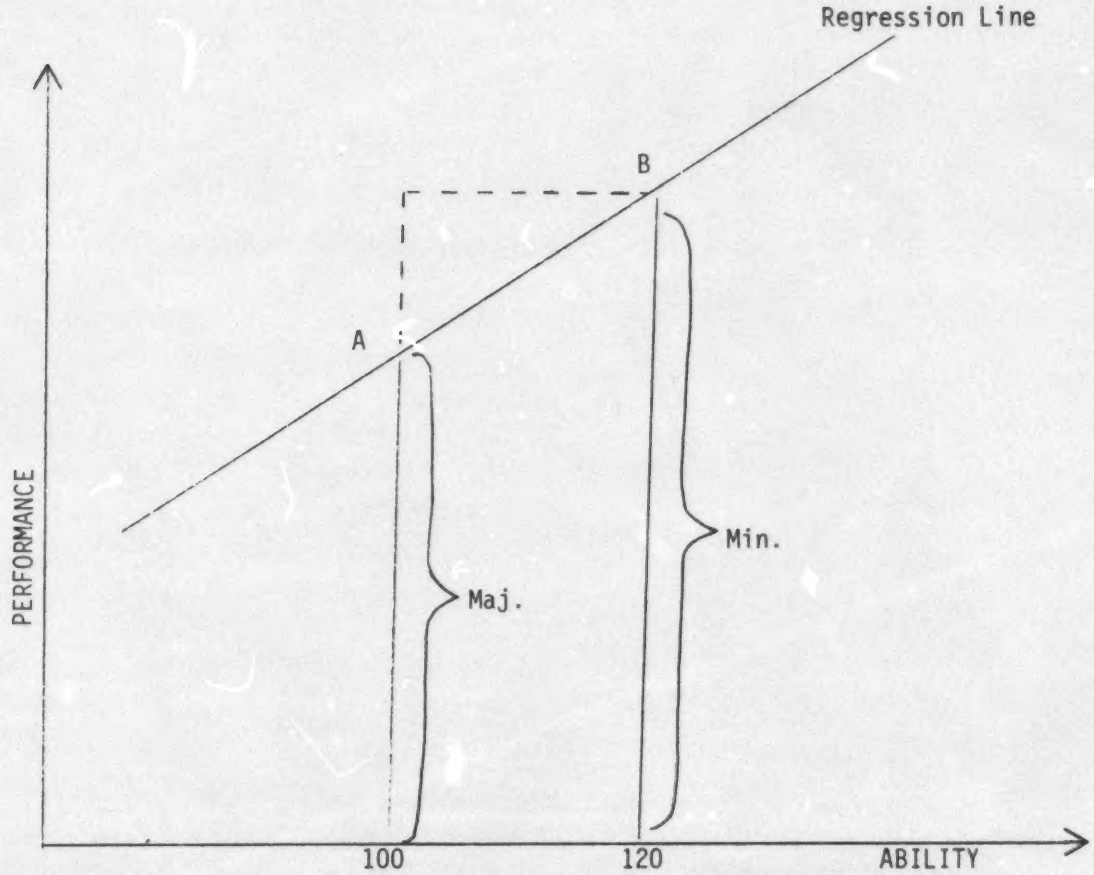
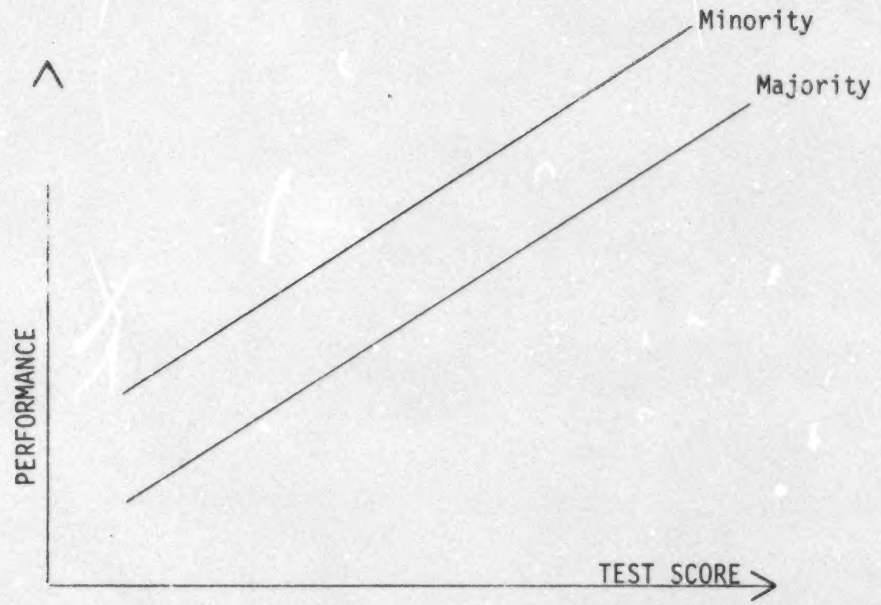


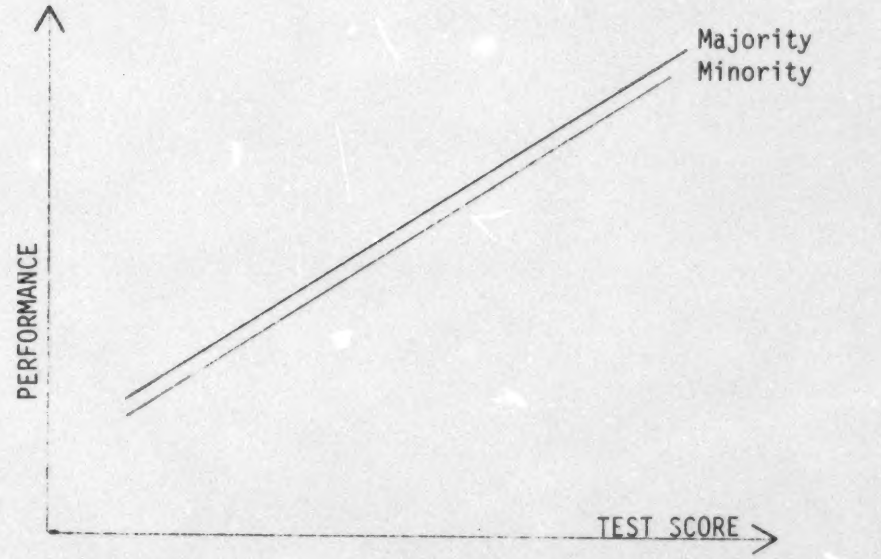
Figure 2. The Hypothesized Relationship Between Ability and Performance for Majority and Minority Applicants if Both Groups All Have the Identical Score of 100 on the Test but the Test is Unfair to Minorities

We will first derive the test-bias prediction about regression lines. Consider a typical validation study. Each worker has two scores: a test score and a job performance score. Plot this pair of scores as a point in a two-dimensional graph such as that shown in Figure 1. The set of such points is called a scatterplot and the tightness of the scatter shows the strength of the relationship between test score and job performance in the study. The points on the scatterplot will not fall perfectly on a line because no test perfectly predicts job performance. However, a number of cumulative studies (see Hunter and Schmidt, in press; Schmidt, Hunter, McKenzie, and Muldrow; 1979 for a review of these studies) have shown that in the employment literature these scatterplots each lie about in a straight line. That straight line is called the regression line of the scatterplot. The regression line is defined in terms of the mean performance of subgroups based on test scores. That is, the value on the regression line above each test score is the mean job performance for all those workers with that test score. Thus, the regression line shows mean performance on the job as a function of test score.

The picture for the test-bias hypothesis is complicated by the fact that there are two regression lines (representing two scatterplots), one for each racial group. We will show the predicted difference between these regression lines by first showing the difference between the mean performance for racial groups at any one test score. Consider two groups of applicants, a majority group and a minority group, all of whom have the same test score, e.g., a score of 100. If the test is valid, then there will be a regression line such as that shown in Figure 2 which relates mean job performance to ability. For the majority group, the test score is an accurate measure of the ability assessed by the test and hence their mean performance will be the value on the regression line immediately above their test score (point A on the regression line). But suppose the tests were unfair to the minority group. Then, although the test score of the minority group members is 100, their actual ability is higher than is measured by the test. For sake of argument, assume that this actual ability score is 120, though the particular numerical value is unimportant. Since the minority-group true ability is 120 rather than 100, their mean job performance will not be the value above 100, but rather the value above their true ability score of 120 (point B on the regression line). As shown in Figure 2, this means that if the two groups of majority and minority were matched at a score of 100, and the mean job performance of each group were plotted separately, then the mean job performance for the minority group would lie above the mean for the majority group. This means if job performance were plotted as a function of test score rather than true ability, then the regression lines for the majority and minority groups would be different, since the value for the minority group would always lie above the value for the majority group.



a. THEORY: The Relationship Between the Regression Lines for Performance on Test Score According to the Hypothesis that Tests are Unfair to Minority Group Members



b. FACTS: A Typical Pair of Regression Lines for Majority and Minority Applicants as Found in a Cumulation of Studies of the Relationship Between Performance and Scores on Cognitive Tests

Figure 3. The Empirical Disconfirmation of the Hypothesis that Cognitive Tests are Unfair to Members of Minority Groups

In Figure 3a the hypothesis of test unfairness is extended from the one score of 100 to the full range of test scores. For each test score, minority group members actually have higher ability than their test scores indicate and hence higher ability than majority group members with the same test scores. Thus, the mean performance at each test score individually would be higher for minority group members than for majority group members, and this would imply that the regression line as a whole for minority group members would lie above the regression line for the majority group.

While Figure 3a represents the hypothesis that tests are unfair to minority group members, Figure 3b shows the empirical results which have emerged from ten years of research. Whereas, those who believed the tests might be unfair to minorities expected minorities to do better on the job than would be predicted by test score (Figure 3a). However, the data show that, where there is a difference, it is a small difference in the opposite direction, i.e., that test scores generally overpredict rather than underpredict the job performance of minority groups (Figure 3b). The studies showing this phenomenon for black applicants are reviewed in Gordon and Rudert (1975); Jensen (1980); Bartlett, Bobko, Mosier, and Hannan (1978); Schmidt and Hunter (1974); Campbell, Crooks, Mahoney, and Rock (1973); Gael and Grant (1972); Gael, Grant, and Richie (1975a); Gael, Grant, and Richie (1975b); Grant and Bray (1970); Ruch (Note 7); and Tenopyr (Note 8). Evidence showing that tests overpredict rather than underpredict the performance of Hispanic workers is reviewed in Schmidt, Pearlman, and Hunter (1980) and Gordon (1975).

If tests were biased against minorities, then they would be biased in all contexts. Thus, the test-bias hypothesis would require that tests used to predict academic achievement would also underpredict achievement. Again findings are exactly the reverse. Reviews of the evidence showing overprediction of academic achievement are Gordon and Rudert (1979), Jensen (1980), Reynolds (in press), and Linn (1975).

The small degree of overprediction of black performance does not mean that tests are slightly biased against whites. The explanation for small amounts of overprediction were given in Linn and Werts (1971). First, they noted that a small portion of overprediction would be predicted by unreliability in the test. Second, they noted that if more than one ability is relevant to the job and if there are racial differences on the other abilities as well, then differences on other abilities will produce differences on each ability considered one at a time. In either case, they predict that the small overprediction is an artifact of considering ability tests one at a time. If composite ability scores across all the relevant ability dimensions are considered, then the overprediction should vanish. A nice example showing just this effect can be found in Powers (Note 6). Powers analyzed data collected from 29 law schools. He found that the extent of overprediction of black achievement dropped from 7.93 to 2.40 to 2.20 points as he moved from the analysis of single predictors to the composite of two predictors to a

composite of three predictors. The overprediction for Hispanic students dropped from 7.50 to 2.90 to 2.60 points as he shifted from single predictors to a composite of all three.

The evidence from all these studies is clear. The regression lines for composite predictors are identical for white, black, and Hispanic workers. There is no underprediction of minority performance. The hypothesis of test bias against minority members is disconfirmed.

Test Fairness and Heredity

There is no single group validity. There is no differential validity. There is no underprediction of minority performance. All the key predictions of the hypothesis of bias against minority test takers have been empirically disconfirmed by thousands of pieces of independent information. Thus, there can be no doubt that tests are fair to minorities.

What then is the meaning of the differences in mean test scores for different ethnic groups? Is this evidence of hereditary differences or could it be the result of cultural disadvantage? These questions cannot be answered on the basis of the evidence cited here. Studies relevant to test fairness start from the point of test administration and predict forward in time. Thus, the finding of test fairness means that a test score represents the person's ability at the time that the test is taken and predicts future events that depend on ability from that point.

On the other hand, questions as to the origins of test differences start from the time of taking the test and work backwards in time. Thus, evidence bearing on heredity must come from different sources. For example, the evidence from twin studies suggests that heredity accounts for, at most, 75 percent of the variance in adult test scores. This suggests that adult ability levels can be far distant from the level of hereditary potential. On the other hand, adult ability levels are very resistant to training (see for example USES, 1970, pp. 275-276). Thus, the environmental effects in ability scores may not be cultural. Noncultural environment factors which may be important include chemical disturbances during embryonic development, trauma, disease and maturational disturbance during childhood, trauma and disease during adult life. Some of these may be related to cultural life style. For example, poor people get poorer prenatal care than rich people. Thus, even if the differences between ethnic groups are due to environmental factors, those factors may not be cultural in the psychological sense.

FAIRNESS OF THE GATB

Differential Validity

For the last 10 years, the U.S. Employment Service has strived to find enough black workers for each validity study to permit separate analysis. At present, 51 such validity studies have been completed. Checks for differential validity and for underprediction of black performance have been run.

The results mirror the results for the field as a whole as reported above; there is no differential validity and tests overpredict rather than underpredict black performance on the job.

The U.S. Employment Service predicts job performance using the GATB (General Aptitude Test Battery) which is scored in terms of nine particular aptitudes (see Table 1). There is a correlation with job performance for each aptitude in each study. Thus, across 51 studies, there are 51(9)=459 opportunities to observe differential validity. In an unpublished study, an Employment Service representative, John Hawk, used computer analysis to test each pair of correlations for significance. Significant differences were found in 31 of the pairs. This is 6.75 percent which differs only trivially from the 5 percent chance level. Since standard deviations for blacks tended to be slightly smaller than standard deviations for whites, slopes were also tested for significance; 30 out of 459 such pairs were significant. This is 6.54 percent which differs even less from the 5 percent chance level. Thus, Hawk found no differential validity in predicting job performance using the GATB.

Table 1

The Extent of Overprediction of Black Job Performance by the GATB

		Partial Correlation	Observed Overprediction	Corrected Overprediction
1a. Overprediction using the specific aptitudes of the GATB.				
General Intelligence	G	.10	.23	.12
Verbal Aptitude	V	.14	.31	.20
Numerical Aptitude	N	.13	.28	.16
Spatial Aptitude	S	.14	.31	.17
Form Perception	P	.14	.30	.19
Clerical Perception	Q	.16	.34	.24
Motor Coordination	K	.19	.39	.38
Finger Dexterity	F	.17	.67	.59
Manual Dexterity	M	.18	.31	.28
1b. Overprediction using the composite general ability scores of the GATB.				
Cognitive Ability	GVN	.11	.25	.18
Perceptual Ability	SPQ	.10	.22	.13
Psychomotor Ability	KFM	.16	.33	.30

Overprediction of Black Performance

If the GATB were unfair to black applicants, then the average job performance of black workers would be higher than is predicted by the white regression line. On the other hand, previous research in other settings suggests that the white regression line actually overpredicts mean black job performance. This can be tested in the U.S. Employment Service in several ways. First, since slopes for blacks and whites are equal, overprediction can be assessed using the partial correlation suggested by Darlington (1971); i.e., $r_{yc \cdot x}$ where x is test score, y is job performance, and c is a dummy variable coded for race using c=+1 for whites and c=-1 for blacks. This partial correlation would be negative if the test were unfair to black workers, and is positive to the extent that the test actually overpredicts black job performance. Second, we can assess the exact amount by which mean job performance for whites is greater than mean job performance for blacks if groups are matched for test score. However, this amount is biased to the extent that the test is not a perfect measure of ability. This can be corrected using the reliability of the test. This leads to a third estimate of overprediction, which is the corrected amount eliminating the bias due to error of measurement.

Table 1a shows the measurement of overprediction using the cumulative statistics for the GATB for each of the nine aptitudes measured by the GATB. Table 1a shows that all nine aptitudes overpredict job performance for blacks, though not to the same extent. The psychomotor aptitudes overpredict black performance to a much greater extent than is true for the cognitive or perceptual aptitudes. The explanation for this lies in the Linn and Werts (1971) observation: Overprediction by one aptitude is caused by group differences on other relevant aptitudes. Hunter (Note 3) has shown that cognitive ability (as measured on the GATB by a composite score for G, V, N or general intelligence, verbal aptitude, and numerical aptitude) is relevant to almost all jobs. Thus, even though differences in psychomotor aptitude are controlled by matching on the most relevant psychomotor aptitude, there will still be differences in performance on the job between black and white workers because they differ in cognitive ability.

Hunter (Note 2, Note 3) has recently shown that the GATB can also be effectively scored for three general abilities instead of nine particular aptitudes. These composite scores are cognitive ability (i.e., GVN, the composite of intelligence, verbal and numerical aptitude), perceptual ability (i.e., SPQ, the composite of spatial aptitude, form perception, and clerical perception), and psychomotor ability (i.e., KFM, the composite of motor coordination, finger dexterity, and manual dexterity). Table 1b shows the extent of overprediction for the composite scores measuring the three general abilities. All indicators show the composite scores to be fair to black workers; all show overprediction of black performance.

The extent of overprediction is smaller for composite scores than for specific aptitudes. For example, the average corrected overprediction is .26 for specific aptitudes and only .20 for general abilities. This reflects the

fact that a general ability controls for more of the relevant difference in job performance than does a specific aptitude.

Work is under way to analyze overprediction in the context of the job families determined by Hunter (Note 3) which are known to have similar ability profiles. Preliminary calculations have already shown that abilities will have only negligible overprediction for jobs to which they are highly relevant and large overprediction for jobs where they have only low relevance. This work may confirm the hypothesis of Linn and Werts (1971) in that it will show that overprediction is an artifact of considering tests one at a time rather than in composites tuned to the job under consideration.

Conclusion

The GATB is fair to minority applicants. Once the correct ability composite is used for a given job, there is no difference in the mean job performance between majority and minority workers with the same composite ability score. That is, there is no difference in mean work performance between different workers if they have the same pattern of aptitudes.

RACIAL AND ETHNIC DIFFERENCES ON THE GATB

Overview

Racial and ethnic groups do not have identical patterns of scores on ability tests. This is true for the GATB as well. These differences were once thought to be impossible, and people hypothesized that the observed differences were the result of racially or ethnically biased items in the tests. This has now been proven false. The differences in ability are real; differences in ability are accompanied by corresponding differences in job performance. These differences may be due to cultural disadvantage, but that does not alter the very real differences in the probability of high job performance between different groups on certain jobs.

If we can define "merit" as hiring that person who is likely to do best at the job, then optimal test use would dictate hiring that person who is highest on the composite ability score relevant to that job. This is called "ranking" in civil service jargon. It is economically optimal in the sense that it maximizes the average job performance of those selected. That is, ranking produces maximal productivity in the hired work force.

However, ranking also guarantees a racial and ethnic imbalance in the work force. The percentage of black and Hispanic workers will normally be less than the percentage of black or Hispanic applicants. This varies from job to job but will be present to some extent in every job. For the GATB, disparities in hiring can be shown to differ from one job family to the next depending on job complexity.

There is another irony in personnel selection. The more selective the employer can be in choosing workers, the higher the average productivity of

the people selected. Thus, the lower the selection ratio, the greater the economic benefit due to the use of ability tests in selection. But the more selective the procedure, the greater the disparity in racial and ethnic hiring rates. Thus, racial imbalance will be at its highest in exactly those jobs where the test is most useful to the employer (and indirectly to those who are served by the employer).

For employers who can afford the economic loss, it is possible to trade-off higher labor costs for racial balance. There are various schemes for achieving racial balance, some of which are more costly than others. The worst method of all is the low-cutoff method in which workers are hired at random from among those who pass a very low cutoff on the test. These schemes are discussed in detail in Hunter (Note 4); Hunter, Schmidt, and Rauschenberger (in press, 1977); and in Cronbach, Yalow, and Scaeffer (1980).

Misinterpretation of Mean Differences

The quantitative analysis of aptitude and ability differences is usually in terms of group means. It is important to note in advance that mean differences between two groups do not imply that people in the group with the higher mean score uniformly higher than people in the group with the lower mean. Rather it should be noted that people in both groups will be found at every level of ability. It is just that the frequency of a given level will vary from group to group. For example, the probability of high ability is lower in the group with the lower mean, and the disparity in frequency is greater for higher ability levels. Similarly, the frequency of low ability is higher in the group with the lower mean and the disparity in frequency is greater for extreme low levels than for moderate levels.

Thus, if there are group differences on an ability test used for selection, then no group will be excluded. There will be members of every group whose ability is higher than the selection cutoff. However, there will be a disparate impact; groups with lower mean scores will have fewer applicants hired.

Mean Differences on the GATB

Means for various groups are reported in the GATB Manual (USES, 1970, pp. 277-288). Table 2a shows the means from the Manual, Table 17-12 for five groups on the GATB aptitudes. Means for these groups on ability composites are shown in Table 2b. One way of converting these means into frequencies was used to construct Tables 2c and 2d. Table 2c shows the percentage of each group that will score higher than the majority group average on each aptitude. This percentage is by definition 50 percent for the majority group. Table 2d shows similar percentages for the three general abilities measured by the GATB. Note that some of the people in each group score above the average for the majority and that in some groups more than half score above the majority average (see Orientals and Mexican Americans on the psychomotor aptitudes for example).

Table 2

Differences in Aptitude and Ability Distributions on the GATB

2a. Mean Aptitude Scores on the GATB (Taken from Table 17-12 in USES, 1970), Aptitudes Defined in Table 1 (SD=20)

	Cognitive			Perceptual			Psychomotor		
	G	V	N	S	P	Q	K	F	M
Majority	102	101	97	106	102	102	100	96	103
Oriental	98	95	98	103	101	105	110	98	108
Mexican American	91	90	88	100	98	97	103	98	107
Black	84	86	83	91	91	94	98	89	100
Indian	84	82	82	103	101	98	105	98	113

2b. Mean Ability Scores on the GATB for the Three General Abilities (with Sample Sizes for the Five Groups)

	Cognitive GVN	Perceptual SPQ	Psychomotor KFM	Sample Size
Majority	300	310	299	6672
Oriental	291	309	315	136
Mexican American	268	295	308	1425
Black	253	276	287	1413
Indian	248	302	312	171

2c. Percentage Who Score Above the Majority Average on Each GATB Aptitude

	Cognitive			Perceptual			Psychomotor		
	G	V	N	S	P	Q	K	F	M
Majority	50	50	50	50	50	50	50	50	50
Oriental	42	38	52	44	48	56	69	54	60
Mexican American	29	29	33	38	42	40	56	54	58
Black	18	23	24	23	29	34	46	36	44
Indian	18	17	23	44	48	42	60	48	69

2d. Percentage Who Score Above the Majority Average on the Three GATB General Abilities

	Cognitive GVN	Perceptual SPQ	Psychomotor KFM
Majority	50	50	50
Oriental	45	49	63
Mexican American	31	39	57
Black	23	25	41
Indian	20	44	61

Table 3

Racial and Ethnic Group Means on the GATB General Abilities Expressed in Majority Group Standard Scores

	Cognitive	Perceptual	Psychomotor
Majority	.00	.00	.00
Oriental	-.13	-.02	+.34
Mexican American	-.51	-.29	+.18
Black	-.75	-.67	-.23
Indian	-.85	-.16	+.27

Table 3 shows the same relationships in a different way. In Table 3 the majority group is used as a baseline and is given a mean of 0. The means for other groups are shown relative to the majority group, with negative means if they are below the majority mean and positive means if they are above the majority mean on each given ability. The unit of measure is chosen to be that of the standard score. This means that 68 percent of each group has scores within one unit of the mean, and has 95 percent of its scores within two units of the mean. By definition, 68 percent of the majority group would have scores between +1 and -1, while 95 percent would lie between +2 and -2. Other groups vary by ability. For example, Orientals have a mean of +.34 on psychomotor ability. Thus, their scores are centered about +.34 rather than 0, with 68 percent between -.66 and +1.34 and with 95 percent between -1.66 and +2.34.

Table 3 makes it clear that groups are separated much more on cognitive ability than on perceptual or psychomotor ability. On psychomotor ability, three of the groups have a higher mean than does the majority.

Job Complexity

Hunter (Note 3) analyzed the 515 validation studies carried out by the U.S. Employment Service in terms of job families. He sought to find job analysis systems which would create families of relatively homogeneous ability requirements. Validation could then be done by job family rather than by single job. A number of job analysis systems proved successful for this purpose. One such system is a set of five categories along a dimension called "complexity." This dimension was created from the Data-People-Things dimensions defined by Fine (1955; Fine and Heinz, 1958) to assess skill and responsibility levels in a job. The basic facts about the complexity job families are given in Table 4.

Table 4

Job Families Determined by Complexity.

4a. The Defining Dimensions and Categories for the Job Complexity Families

Level	Dimension	Categories
1	Things	Set up work
2	Data	Synthesize/Coordinate
3	Data	Analyze/Compile/Compute
4	Data	Compare/Copy
5	Things	Feeding/Offbearing

4b. The Multiple Regression Equation for Predicting Job Performance in Each Complexity Category (EJP = Estimated Job Performance, GVN = Cognitive Ability, SPQ = Perceptual Ability, KFM = Psychomotor Ability)

Level	Equation	Multiple Correlation
1	EJP = .40 GVN + .19 SPQ + .07 KFM	.59
2	EJP = .58 GVN	.58
3	EJP = .45 GVN + .16 KFM	.53
4	EJP = .28 GVN + .33 KFM	.50
5	EJP = .07 GVN + .46 KFM	.49

4c. The Group Means on Estimated Job Performance in Majority Group Standard Scores

	Complexity Level				
	1	2	3	4	5
Majority	.00	.00	.00	.00	.00
Oriental	-.05	-.13	-.01	+.15	+.30
Mexican American	-.42	-.51	-.38	-.17	+.10
Black	-.75	-.75	-.71	-.57	-.32
Indian	-.60	-.85	-.64	-.30	+.13

Table 4a shows the definition of each complexity category in terms of the Data-People-Things dimensions. The first and last categories are defined in terms of the Things dimension. These are also the smallest categories in terms of number of jobs. Basically, the complexity dimension can be regarded as Fine's Data dimension with the two Things categories pulled out for

special consideration. The People dimension proved useless since nearly all jobs fell in one category.

Table 4b shows the regression equation that predicts job performance in each complexity category. As complexity goes down the relevance of cognitive ability decreases, while the relevance of psychomotor ability increases. Industrial setup work is the only job category in this system which has a special contribution from perceptual ability; but work is now underway to locate other such particular job classes. The multiple correlation assessing the extent of prediction of job performance decreases as job complexity decreases. This may reflect a tendency for individual differences to be more constrained in lower level jobs.

Table 4c shows the mean estimated job performance for each racial and ethnic group at jobs of each level of complexity. The rank order changes as job complexity decreases. The majority group has the highest mean performance in the top three categories but ranks second on the fourth category and only fourth in the last category. Fairness studies have shown that the estimated job performance means shown in Table 4c are mirrored in actual job performance means if hiring is random.

Given the real differences in ability and job performance among racial and ethnic groups, it follows that optimal personnel selection will result in different hiring rates for different groups in different jobs. Since ability differences vary in different ways for different jobs, differences in hiring rates will vary in different ways for different jobs. Thus, hiring rates will be shown for each job complexity category separately.

The largest determinant of hiring rates is the selection ratio. The selection ratio is the ratio of the number of persons hired to the number of applicants. The smaller this number, the more selective the hiring can be. As the selection ratio decreases, hiring rates will decrease for all groups, but not proportionately. The disparities between ethnic groups become relatively larger as the selection ratio decreases.

In the fiscal year 1979-80 the U.S. Employment Services tried to find employment for 17,974,684 persons. It found temporary jobs of 3 days or less for 365,502 persons, temporary to seasonal jobs of 4-150 days for 1,141,766 persons, and permanent jobs for 2,460,156 persons. If we consider only the permanent jobs, then the selection ratio for employers using the Employment Service is 2,460,156/17,974,684 = 13.69 percent. Hunter (Note 4) has shown this to be a very selective ratio which generates very large savings in labor costs for employers who use optimal selection.

An application population which is a mixture of racial and ethnic groups does not have a truly normal distribution. Thus, the determination of the cutoff score which provides a given overall selection ratio such as 13.69 percent is a complicated computation (see the appendix of Hunter, Schmidt, and Rauschenberger, 1977). Furthermore, the cutoff score varies with the mix of

groups and hence varies from office to office. Therefore, tables have been computed in terms of the selection ratio for the majority group rather than the overall selection ratio. That is, if the selection ratio is listed as 50 percent, then 50 percent of the majority group will be selected. The overall selection ratio will vary depending on the population mix and depending on the kind of job.

Table 5

Percentage of Each Ethnic or Racial Group Selected
as a Function of Selection Ratio and Job Complexity Family

	Selection Ratio				
	50	35	15	10	5
Complexity Level 1					
Majority	50	35	15	10	5
Oriental	48	33	14	9	5
Mexican American	34	21	7	4	2
Black	27	17	5	3	1
Indian	23	13	4	2	1
Complexity Level 2					
Majority	50	35	15	10	5
Oriental	45	30	12	8	4
Mexican American	31	18	6	4	2
Black	23	13	4	2	1
Indian	20	11	3	2	1
Complexity Level 3					
Majority	50	35	15	10	5
Oriental	50	34	15	10	5
Mexican American	35	22	8	5	2
Black	26	15	5	3	1
Indian	24	14	4	2	1
Complexity Level 4					
Oriental	56	41	19	13	7
Majority	50	35	15	10	5
Mexican American	43	29	11	7	3
Indian	38	25	9	6	3
Black	28	17	5	3	1
Complexity Level 5					
Oriental	62	46	23	16	9
Indian	55	40	18	12	7
Mexican American	54	39	17	12	6
Majority	50	35	15	10	5
Black	37	24	9	5	3

Table 5 presents the hiring rates for different racial and ethnic groups as a function of selection ratio within each level of job complexity. Note that the rank order of groups differs in complexity level. The U. S. Employment Service selection ratio of 13.69 percent is represented in this table by the level of 15 percent for the majority group. For Complexity Levels 1 and 2 the majority has a higher hiring rate than any of the minority groups. For Complexity Level 3 the hiring rate for Orientals is as high as that for the majority. For Complexity Level 4 the hiring rate for Orientals is higher than that for the majority. For Complexity Level 5 the hiring rates of Orientals, Indians, and Mexican Americans are all higher than the hiring rate for the majority. The only group with a lower hiring rate in all job categories is the black group and it is also the only group with lower ability means on all aptitudes.

Conclusion

Racial and ethnic groups differ in their distribution of ability. Groups differ in terms of the number of people with very high ability on any given aptitude. Thus, if people are hired on the basis of ability then the percentage of persons from a given group will vary. To the extent that a given ability is relevant to a given job, then groups low on that ability will have a lower hiring rate than groups with high means. For high complexity jobs (Levels 1-3) only Orientals have hiring rates approaching those of the majority. However, for jobs of low complexity, minority hiring rates for some groups exceed the hiring rate for the majority. For Level 5 the hiring rates for three out of four minority rates exceed the rate for the majority.

The finding that minority hiring rates exceed those for the majority in certain jobs is a considerable departure from the contemporary personnel literature. This stems from the fact that the GATB is unique in using psychomotor ability to predict job proficiency. No other major battery includes psychomotor aptitudes. For jobs of low complexity, the use of psychomotor ability simultaneously lowers adverse impact while increasing validity.

REFERENCE NOTES

1. Hunter, J. E. An analysis of validity, differential validity, test fairness, and utility for the Philadelphia Police Officers Selection Examination prepared by the Educational Testing Service. Report to the Philadelphia Federal District Court, Alvarex vs. City of Philadelphia, 1979.

2. Hunter, J. E. The dimensionality of the General Aptitude Test Battery (GATB) and the dominance of general factors over specific factors in the prediction of job performance. U. S. Employment Service, 1982.

3. Hunter, J. E. Test validation for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB). Report to the U. S. Employment Service, 1982.

4. Hunter, J. E. The economic benefits of personnel selection using ability tests: A state of the art review including a detailed analysis of the dollar benefit of U. S. Employment Service placements and a critique of the low-cutoff method of test use. Report to the U.S. Employment Service, 1981.

5. Mack, M. J., Schmidt, F. L., and Hunter, J. E. Dollar implications of alternative models of selection: A case study of park rangers. Unpublished manuscript available through Frank L. Schmidt, Office of Personnel Management, Washington, D.C. 20415.

6. Powers, D. E. Comparing predictions of law school performance for black, Chicano, and white students. Law School Academic Council, LSAC-77-3, 1977.

7. Ruch, W. W. A re-analysis of published differential validity studies. Paper presented at the symposium "Differential validation under EEOC and OFCC testing and selection regulations." American Psychological Association, Honolulu, Hawaii, 1972.

8. Tenopyr, M. L. Race and socio-economic status as moderators in predicting machine-shop training success. Paper presented at the American Psychological Association; Washington, D.C., 1967.

REFERENCES

Bartlett, C. J., Bobko, P., Mosier, S. B., and Hannon, R. Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. Personnel Psychology, 1978, 31, 233-241.

Boehm, V. R. Differential predicition: A methodological artifact? Journal of Applied Psychology, 1977, 62, 146-154.

Campbell, J. T., Crooks, L. A., Mahoney, M. H., and Rock, D. A. An investigation of sources of bias in the prediction of job performance. A six year study. Final Project Report No. PR-73-37. Princeton, N. J.: Educational Testing Service, 1973.

Cronbach, L. J., Yalow, E., and Schaeffer, G. A mathematical structure for analyzing fairness in selection. Personnel Psychology, 1980, 33, 693-704.

Darlington, R. B. Another look at "culture fairness." Journal of Educational Measurement, 1971, 8, 71-82.

Fine, S. A. A structure of worker functions. Personnel and Guidance Journal, 1955, 34, 66-73.

Fine, S. A. & Heinz, C. A. The functional occupational structure. Personnel and Guidance Journal, 1958, 37,180-192.

Gael, S., and Grant, D. L. Employment test validation for minority and nonminority telephone company service representatives. Journal of Applied Psychology, 1972, 56, 135-139.

Gael, S., Grant, D. L., and Ritchie, R. J. Employment test validation for minority and nonminority clerks with work sample criteria. Journal of Applied Psychology, 1975, 60, 420-426(a).

Gael, S., Grant, D. L., and Ritchie, R. J. Employment test validation for minority and nonminority telephone operators. Journal of Applied Psychology, 1975, 60, 411-419(b).

Ghiselli, E. E. The validity of aptitude tests in personnel selection. Personnel Psychology, 1973, 26, 461-477.

Gordon, R. A. Examining labeling theory: The case of mental retardation. Pp 83-146 in Gove, E. R. (Ed.) The labeling of deviance: A perspective. Beverly Hills, Cal.: Sage/Halstead, 1975.

Gordon, R. A., and Rudert, E. E. Bad news concerning IQ tests. Sociology of Education, 1979, 52, 174-190.

Grant, D. L., and Bray, D. W. Validation of employment tests for telephone company installation and repair occupations. Journal of Applied Psychology, 1970, 54, 7-14.

Humphreys, L. G. Statistical definitions of test validity for minority groups. Journal of Applied Psychology, 1973, 58, 1-4.

Hunter, J. E., and Schmidt, F. L. Fitting people to jobs: Implications of personnel selection for national productivity. Chapter to appear in E. A. Fleishman (Ed.), Human Performance and Productivity, in press.

Hunter, J. E., and Schmidt, F. L. Differential and single group validity of employment rests by race: A critical analysis of three recent studies. Jounral of Applied Psychology, 1978, 63, 1-11.

Hunter, J. E., and Schmidt, F. L. A critical analysis of the statistical and ethnical implications of five definitions of test fairness. Psychological Bulletin, 1976, 83, 6, 1053-1071.

Hunter, J. E., and Schmidt, F. L. and Hunter, R. F. Differential validity of employment tests by race. A comprehensive review and analysis. Psychological Bulletin, 1979, 86, 721-735.

Hunter, J. E., Schmidt, F. L., and Rauschenberger, J. Methodological and statistical issues in the study of bias in mental testing. Chapter in Reynolds, C. R. and Brown, R. T. (Eds.) Perspectives on bias in mental testing. New York: Plenum Press, in press.

Hunter, J. E., Schmidt, F. L., and Rauschenberger, J. M. Fairness of psychological tests: Implications of four definitions for selection utility and minority hiring. Journal of Applied Psychology, 1977, 62, 245-260.

Jensen, A. R. Bias in mental testing. New York: Free Press, 1980.

Katzell, R. A. and Dyer, F. J. Differential validity revived. Journal of Applied Psychology, 1977, 62, 137-145.

Linn, R. L. Test bias and the prediction of grades in law school. Journal of Legal Education, 1975, 27, 293-323.

Linn, R. L. and Werts, C. E. Considerations for studies of test bias. Journal of Educational Measurement, 1971, 8, 1-4.

O'Conner, E. J., Wexley, K. N., and Alexander, R. A. Single-group validity: Fact of fallacy? Journal of Applied Psychology, 1975, 60, 352-355.

Pearlman, K. The validity of tests used to select clerical personnel: A comprehensive summary and evaluation. U. S. Office of Personnel Management, Personnel Research and Development Center, TS-79-1, August, 1979.

Pearlman, K., Schmidt, F. L., and Hunter, J. E. Validity generalization results for tests used to predict success and job proficiency in clerical evaluations. Journal of Applied Psychology, 1980, 65, 103-106.

Reynolds, C. R. The problem of bias in psychological assessment. Chapter in Reynolds, C. R. and Gutkin, T. B. (Eds.) A handbook for school psychology. New York: John Wiley and Sons, in press.

Rosenfeld, M., and Thornton, R. F. The development and validation of a multi-jurisdictional police examination. Princeton, N. J.: Educational Testing Service, 1976.

Schmidt, F. L., and Hunter, J. E. Racial and ethnical bias in psychological tests: Divergent implications of two definitions of test bias. American Psychologist, 1974, 29, 1-8.

Schmidt, F. L., Berner, J. G., and Hunter, J. E. Racial differences in validity of employment tests: Reality or illusion? Journal of Applied Psychology, 1973, 53, 5-9.

Schmidt, F. L., Hunter, J. E., McKenzie, and Muldrow, T. The impact of valid selection procedures on workforce productivity. Journal of Applied Psychology, 1979, 64, 609-626(b).

Schmidt, F. L., Hunter, J. E., Pearlman, K., and Shane, G. S. Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. Personnel Psychology, 1979, 32, 257-281(a).

Schmidt, F. L., Pearlman, K., and Hunter, J. E. The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. Personnel Psychology, 1980, 33, 705-724.

U.S. Employment Service. Section III of the Manual for the USES General Aptitude Test Battery. U. S. Department of Labor, 1970.

☆ U.S. Government Printing Office: 1983—417-794/1395